

Pre-processing

After the [Collecting the data](#), we had around 650 chats in different languages but no idea which chat was in which language. Furthermore, we had given a promise to anonymize the data and we did not have a tool to browse the data in the available format. Thus, before making the data available to the research team, we had to pre-process them. We thus had to perform some steps before our research could start.

Anonymization

The data in the corpus was anonymized by means with the same methodology that we already applied successfully in the [SMS corpus](#).

General privacy

While the project did not have the intention of collecting private information about the informants (other than what they provided in the questionnaire), it could still not be assumed that the informants would not send personal information in their SMS, so it was the team's task to remove specific pieces of information again. These steps were performed by means of computational linguistics. The stories told, however might still allow you to recognize individual informants. If that is the case, we asked to comply with common [research ethics](#) and keep that knowledge to yourself.

First names

A reference list of first names in different languages was used to remove all first names. As always with such a task, it was a balance act between precision and recall. On the one hand, all first names should be removed from the data, on the other hand no information that is homograph to a first name should get lost. To get the best possible result, it was decided to not actually remove first names, but to rotate them, meaning that the name Peter in an SMS would not get replaced by e.g. [FirstName], but by e.g. Ferdinand. This procedure has several advantages:

1. The text remains easy to read.
2. Because Peter is always replaced with Ferdinand, all occurrences of the same name remain the same. Conversations can therefore easier be recognized as such.
3. Names that did not get replaced because of homography are not recognizable as such, i.e. if the name Minna appears in an SMS, nobody can know, whether this is a replaced name or whether it is a name that was not replaced because it is a homograph to some rare word in Romansh. The scientist working with the data will therefore always assume that first names he comes across are actually not the real names used in the SMS.

Tests show, that more than 95% of all first names were in fact removed.

Lastnames

Only very few last names can in fact be found in the data. Because of this limitation, the same procedure as with first names could not be applied, because additionally some of the last names used are very rare if not unique. It was therefore decided to replace all last names with [LastName] instead. In a combined effort of manually analyzing and means of computer linguistics, more than 95% of all last names were removed. Numbers In an effort to remove information about phone numbers, bank

accounts etc., all numbers with three and more digits were removed and each digit was replaced with one N. The phone number 079 987 65 43 would thus become NNN NNN 65 43, while 0799876543 would be NNNNNNNNNN. Reliability here lies with 100%.

E-Mail addresses

All email addresses were removed and replaced with xxx@yyy.ch, while keeping the number of characters. info@uzh.ch would therefore become xxxx@yyy.ch, while admin@google.com would become xxxxx@yyyyyy.com.

Street addresses

Street addresses were removed and replaced by [StreetAddress].

WWW addresses

WWW addresses were kept since they contain information publicly available.

City names

Names of cities were kept because they cannot be considered as private information and because they may be important for the understanding of the text.

From:
<https://whatsup.linguistik.uzh.ch/> -

Permanent link:
https://whatsup.linguistik.uzh.ch/01_corpus/02_preprocessing?rev=1569239471

Last update: **2022/06/27 09:21**

